Soft Computing manuscript No.
(will be inserted by the editor)

Towards Cyberbullying Free Social Media in Smart Cities: A Unified Multi-modal Approach

Kirti Kumari · Jyoti Prakash Singh · Yogesh Kumar Dwivedi · Nripendra Pratap Rana

Received: date / Accepted: date

Abstract Smart cities are shifting the presence of people from physical world to Cyber world (Cyber space). Along with the facilities for societies, the troubles of physical world such as bullying, aggression, hate speech etc., are also taking their presence in Cyber space. This paper aims to dig the posts of social media to identify the bullying comments containing text as well as images. In this article, we have proposed a unified representation of text and images together to eliminate the need for separate learning modules for images and text. A single layer Convolutional Neural Network model is used with the unified representation. The major findings of this research are that text represented as images is a better model to encode the information. We also found a single layer Convolutional Neural Network is giving better results with two-dimensional representation with 2048 filters of one-gram Term Frequency-Inverse Document Frequency only. In the current scenario, we have used three layers of text and three layers of a colour image to represent the input which gives a recall of 74% of the bullying class with one layer of Convolutional Neural Network.

Keywords Online Social Networks · Deep Learning · Convolutional Neural Network · Cyberbullying · TF-IDF

Kirti Kumari
National Institute of Technology Patna, India
E-mail: kirti.cse15@nitp.ac.in

Jyoti Prakash Singh
National Institute of Technology Patna, India
E-mail: jps@nitp.ac.in

Yogesh Kumar Dwivedi
School of management, Swansea University, UK
E-mail: ykdwivedi@gmail.com

Nripendra Pratap Rana
School of management, University of Bradford, UK
E-mail: nrananp@gmail.com

1 Introduction

The smart city is defined as a city that makes optimal use of all the interconnected information available today to better understand and control its operations and optimize the use of limited resources (defined by International Business Machines). Smart city technology can assist towns to function more effectively with the benefits of data-driven decision-making (Visvizi et al., 2018), enhanced citizen and government engagement (Visvizi and Lytras, 2018), safer communication (Lytras and Visvizi, 2018), and improved transportation. It also facilitates flexible, decentralized and intelligent systems for learning (Lytras et al., 2018). Smart city services have shifted the presence of people from physical worlds to the virtual world (cyberspace), e.g. online banking operations, online shopping, online ticket bookings, medical services through telemedicine, etc. Online content is a vital asset of smart city (Alkhammash et al., 2019) and managing online content is critical challenge of today's societies to make sustainable (Visvizi et al., 2019). Along with the facilities for mankind, the troubles of the physical world is also shifted to the cyber world. A good example may be bullying which used to occur in physical worlds has shifted to cyberspace through Online Social Network (OSN) platforms such as Facebook[1], Twitter[2], Instagram[3], YouTube[4], and Reddit[5], etc. OSNs are a platform which offers communication opportunities, gives users a place to engage in social interaction, offer the possibilities of relationships and maintain existing friendships. OSN facilitates social interactions (Torres-Ruiz and Lytras, 2016) by the way of text messaging, posting images, videos, and a combination of these (Steiner-Correa et al., 2018). Along with these benefits, these sites are becoming a stupendous place for the people mainly teenagers and youngsters to harass, threaten, and embarrass others. One of the major issues of concern are Cyberstalking (League, 2011), Cyber-aggression (Chatzakou et al., 2017; Kumari et al., 2019), Cyberbullying (Salawu et al., 2017) and so on. Among them, Cyberbullying is growing fast and becoming a serious problem for sustainable development of today's society (Hosseinmardi et al., 2015; Kumari et al., 2019). Cyberbullying typically refers to repeated and hostile behavior (e.g., hurtful comments, videos, and images) performed to intentionally harass or harm individuals. As social media is a heterogeneous platform, Cyberbullying could occur in various form like, written messages (e-mails, instant messaging, chats, blogs), verbal over phone, visual (posting, sending or sharing embarrassing images or video), exclusion (purposefully excluding someone from an online group), and impersonation (stealing and revealing personal information, using another person's name and account) (Dadvar et al., 2014).

The victims of Cyberbullying are found to suffer from hopelessness, worthlessness, frustration, depression, anxiety, sleep-related issues and, in extreme cases committing suicide (Bhat et al., 2017). Recent studies (Pater et al., 2015) have shown that teenagers make enormous use of image and video sharing online sites like Instagram, and Vine, etc., to share their status. Visual (image and video)

[1] www.facebook.com
[2] www.twitter.com
[3] www.instagram.com
[4] https://www.youtube.com/
[5] https://www.reddit.com

content now accounts for over 70% of all web traffic[6]. A substantial increase in Cyberbullying cases using image and video content has been reported recently (Seiler and Navarro, 2014) which is growing larger and meaner with pictures and video (Kornblum, 2008). The seriousness of the issue requires instant attention from a technical perspective because manual detection is not scalable as well as time-consuming. Automated tools need to be created that can help to minimize potential tragedies in social media and provides an automated surveillance that (Van Royen et al., 2015; Sui et al., 2017; Chui et al., 2018) in a smart city. Most of earlier works (Dadvar and De Jong, 2012; Dinakar et al., 2012; Al-garadi et al., 2016; Badjatiya et al., 2017; Chatzakou et al., 2017; Zhao and Mao, 2016; Davidson et al., 2017) considered only the cases of Cyberbullying in text based post. The other critical information included in the post such as image, audio, video, URLs, etc, were ignored in earlier researches. Recently, Hosseinmardi et al. (2015); Singh et al. (2017) included the image part also in their Cyberbullying detection models but they considered the text part as the main indicative point of bullying. Six possible combination of text and images of a social media post may represent bullying and non-bullying instances.

- Case 1: The text, as well as image, are bullying and the together post is also bullying as shown in Figure 1.
- Case 2: The text is bullying and the image is non-bullying but the together post is bullying as shown in Figure 2.
- Case 3: Both the text and image separately are non-bullying but together it has bullying sense as shown in Figure 3
- Case 4: Neither the text is bullying, nor the image and together they are not bullying.
- Case 5: The text is non-bullying and the image is bullying but together the post has non-bullying sense.
- Case 6: The text is non-bullying and the image is bullying but together the post has bullying sense.

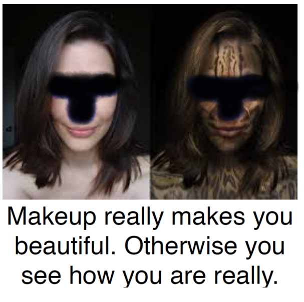

Fig. 1 Cyberbullying Post with having both image and comment are bullying

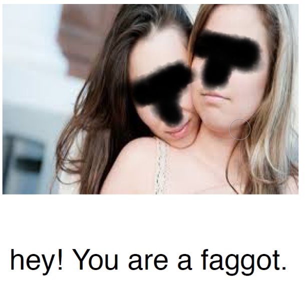

Fig. 2 Cyberbullying Post with having non-bullying image and bullying comment

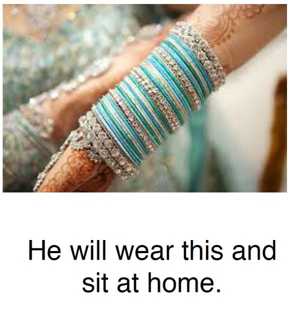

Fig. 3 Cyberbullying Post with having both image and comment are non-bullying

[6] https://www.recode.net/2015/12/7/11621218/streaming-video-now-accounts-for-70-percent-of-broadband-usage

Most of the existing systems employed separate learning module for text and images which trained them independently. These system may never identify the 3^{rd} case of Cyberbullying listed above.

We motivated us to investigate the cases of Cyberbullying using text and images. We developed a model to identify all cases of Cyberbullying with text and image combination. Our main emphasis was on to differentiate and identify the cases of Cyberbullying where both text as well as image separately may look innocent.

We develop a multi-modal deep learning-based system that can be trained on image and related textual comments together to identify the bullying post. For that, we have proposed an unified representation (or embedding) of text and image as $M \times N \times 6$ sized multi-dimensional array. Each image is represented in $M \times N \times 3$ matrix, where M, N, and *3* are width, height, and channel of the image respectively. Similarly, each comment is also represented as $M \times N \times 3$ using Term Frequency-Inverse Document Frequency (TF-IDF).

To the best of our knowledge, no dataset containing heterogeneous post (image and text) is publicly available because most of the bullying posts are deleted as they are detected (Chatzakou et al., 2017). The scarcity of heterogeneous Cyberbullying labelled dataset motivated us to create one. We created a dataset by crawling images from Instagram, Facebook, Twitter, and Google searches by giving a query such as bullying, animal, and ugly images and so on. We manually labelled the dataset into bullying and non-bullying posts. Our dataset will be available on request though the author's mail-id[7].

The main contributions of the research can be summarized as:

- Proposed a novel integrated representation of image and text together to learn the visual and textual patterns of social media posts.
- Proposed a multi-layered Convolutional Neural Network (CNN) model that takes the integrated representation of image and text as input and classifies them into bullying or non-bullying.
- Created a dataset of Cyberbullying posts containing image and associated comments.

The rest of the document is structured as follows. The associated literature is described briefly in Section 2. Section 3 presents our suggested framework for Cyberbullying detection. Section 4 presents the findings of the suggested approach. Section 5 includes discussions on the results and consequences of the present studies. Finally, in Section 6, we conclude the paper.

2 Related Works

Cyberbullying falls within the range of adverse Internet conduct, various varieties of which have been researched in the literature, such as hate speech (Badjatiya et al., 2017), Cyber-aggression (Chatzakou et al., 2017), trolling (Paavola et al., 2016), online harassment (Jones et al., 2013), offensive language (Chen et al., 2012), etc. In this section, we discuss some of the potential works proposed in the automated detection of the Cyberbullying domain. We categorize this section in

[7] kirtics518@gmail.com

two subsections based on the content of the post used to detect Cyberbullying events. (i) Cyberbullying detection considering text only, and (ii) Cyberbullying detection with image and text.

2.1 Text-Based Cyberbullying Detection

One of the early works to detect the Cyberbullying events dealing with harassing on social media was proposed by Yin et al. (2009). They used the dataset from Kongregate, MySpace, and SlashDot to train a Support Vector Machine (SVM) classifier by using local, sentiment, and contextual features. The local and sentiment features were derived by calculating the Term Frequency-Inverse Document Frequency (TF-IDF) of each distinct word whereas, for contextual features, they calculated the average cosine similarity of neighbor posts. The best result reported by them was a F1-score of 0.44 with a Kongregate dataset. They however considered the content of the post only to detect whether a post was related to harassing or not and also the accuracy reported work was very low. Reynolds et al. (2011) did a Cyberbullying detection on the dataset of Formspring social networking site. They used a decision tree as a classifier and found an accuracy of 0.78. The model was tested with a very small dataset of containing 10 user's posts and considered only curse words. Dadvar and De Jong (2012) considered gender as their main features and separately classified bullying posts for male and female on MySpace dataset. They used feature to train the SVM classifier and reported the F1-score of 0.08 and 0.28 for female and male-specific posts respectively. Dinakar et al. (2012) considered indirect bullying posts on the dataset of Formspring and Youtube. However, they restricted their model to identify a subset of Cyberbullying cases of Lesbian, Gay, Bisexual, and Transgender (LGBT) type only. They ignored other types of Cyberbullying such as race, culture, intelligence, physical appearance, social rejection, etc. Their model achieved the F1-score of 0.63 using the SVM classifier. Nahar et al. (2013) incorporated the weighted TF-IDF feature and built a weighted directed graph-based model between two classes of users such as victim and bully. Their work achieved the F1-score of 0.92. Dadvar et al. (2014) used a hybrid approach combining the expert system and machine learning approach for detecting the Cyberbullying cases on YouTube dataset. By using three sets of features: content, activity, and user feature. They found the best result for the hybrid approach with Area Under the Curve (AUC) value of 0.76. Al-garadi et al. (2016) detected Cyberbullying cases from Twitter posts using four sets of features: content, activity, user, and network features with Naive Bayes, SVM, Random Forest, and K-Nearest Neighbors classifiers. Their best result was a recall value of 0.71 with a Random Forest classifier. Chen et al. (2017) detected online harassment using different datasets collected from Twitter, YouTube, MySpace, Formspring, Kongregate, and SlashDot using Naive Bayes, and SVM classifiers. They got the maximum recall value of 0.78 for MySpace dataset. Waseem and Hovy (2016) proposed a model to detected Hate Speech related to Racist and Sexist tweets. They used the character n-grams feature to achieve the F1 score of 0.74. Davidson et al. (2017) also focused on Hate Speech related to Racist, Sexism, and Homophobic tweets detection. Their best results were having precision and recall values of 0.44 and 0.61 respectively. Burnap and Williams (2015) detected Cyberhate on twitter using voted ensemble learning based on Logistic Regression, Random

Forest, Decision Tree, and SVM classifiers. Their best result was reported as F1-score with 0.77 using n-gram features. Bohra et al. (2018) identified Hate tweets on Hindi-English code-mixed tweets using SVM and Random Forest as classifiers, and character n-gram, word n-gram, the particular set of words, and exclamation marks as features. They achieved an accuracy of 0.72 for the character n-gram feature with SVM classifier.

2.2 Image and Text-Based Cyberbullying Detection

Hosseinmardi et al. (2016) considered visual features and utilized the user data such as image, its caption, number of followers, and followee to predict whether a post is Cyberbullying or not. However, their finding was that visual features were not very helpful in Cyberbullying detection. Singh et al. (2017) used both textual and visual features to differentiate Cyberbullying versus Non-Cyberbullying. Their training sample was very less and contained high negative words. Without mentioning negative words in the comment part bullying can be possible where image and comments individually cannot be bullying but together may have a bullying case as shown in our Introduction section. In the heterogeneous form of Cyberbullying detection, the main challenge is to collect, label and process the different forms of information (Wang et al., 2017).

Some of the potential works for Cyberbullying detection are listed in Table 1. Identifying Cyberbullying on social media is a very challenging task due to several reasons like the heterogeneous form of the post (text, image, audio, and video), the improper writing style of online users, and multi-lingual text. One of the main problems in the process of automatic Cyberbullying identification with many modalities of posts is that the complex combination of multiple modalities may not be compatible with each other to make the prediction accurate (Chatzakou et al., 2017; Tommasel et al., 2018; Ali and Angelov, 2018). Additionally, multilingual text and non-standard abbreviation on social media posts make difficult to extract the linguistic features using Natural Language Processing tools. In this work, we have tried to combine the textual and visual features to make a unified representation. This unified representation has been used to train a model to identify bullying posts involving text and images.

3 Methodology

We first describe data collected for this study followed by a high-level block diagram of the proposed model shown in Figure 9. The proposed system consists of four phases. They are (i) Input preparation, (ii) Embedding layer, (iii) Convolutional layer, and (iv) Output layer. Each of them is explained in the following subsections.

3.1 Data Collection and Labelling

We gathered bullying images from three popular OSN platforms Facebook, Twitter, and Instagram by specifying several keywords (or queries) such as ugly, fat,

Table 1 Summary of potential works for Cyberbullying detection

Author	Data	Method	Advantages	Disadvantages
Yin et al. (2009)	Text	SVM	Incorporated sentiment and contextual information	Considered only content based features
Reynolds et al. (2011)	Text	Decision Tree	Considered weighted average of cuss words	Training sample is very small and not considered the context information
Dadvar and De Jong (2012)	Text	SVM	Considered age and gender as user features	Reported results are very poor
Dinakar et al. (2012)	Text	SVM	Considered divergent type of Cyberbullying	Restricted to LGBT type
Nahar et al. (2013)	Text	SVM	Reported results are good	Limited to static set of bully words
Dadvar et al. (2014)	Text	Naive Bayes, SVM, Decision Tree, and Expert System	Incorporated user's activity features	Not considered the location and time of the activity
Al-garadi et al. (2016)	Text	Naive Bayes, SVM, Random forest, and K-Nearest Neighbors	Incorporated network features	Restricted to word used in the tweet not the context
Chen et al. (2017)	Text	SVM, and Naive Bayes	Worked with multiple social media platforms	Considered only content based features
Waseem and Hovy (2016)	Text	Logistic Regression	Incorporated linguistic and character n-gram features	Not considered user's gender and location information
Davidson et al. (2017)	Text	Logistic Regression, Naive Bayes, SVM, Decision Tree, and Random forest	Focused on homophobic slurs	Poor difference between Hate and Offensive classes
Burnap and Williams (2015)	Text	Logistic Regression, Random Forest, Decision Tree, and SVM	Focused on antagonistic content	X
Bohra et al. (2018)	Text	SVM, and Random Forest	Multi-lingual consideration	Considered only Hindi and English languages
Hosseinmardi et al. (2016)	Image and text	Logistic Regression	Considered visual features	Works very poor without negative words in social media post
Singh et al. (2017)	Image and text	Bagging Classifier	Addition of visual features into predictive of Cyberbullying	Training samples are very small and works only with negative words in the comment

animal, cartoons of human, porn images and so on. We also used the Google search as a source for searching images by specifying the same queries. A total of 2100 images were collected from these (Facebook, Twitter, Instagram, and Google) sources. To the best of our knowledge, no data for the raised issue is available publicly. So, we had to create own dataset to train and test the proposed model. Since the comment part was not available for all the images. So, we asked seven of our undergraduate students to write a comment for each image. After commenting, each data had two fields, an image, and its comment. Closer observations revealed that there are six possible cases of bullying and non-bullying arises in social media:

- Case 1: where image and comments both are bullying and together it also has bullying sense (Figure 1).
- Case 2: where the image is non-bullying and comment is bullying and together it has bullying sense (Figure 5).
- Case 3: where both image and comment are non-bullying but together it has bullying sense (Figure 4).
- Case 4: where image and comments both are non-bullying and together also is non-bullying (Figure 6).
- Case 5: where the image is bullying and comment is non-bullying together it has a non-bullying sense (Figure 8).

– Case 6: where the image is bullying and comment is non-bullying but together it has bullying sense (Figure 7).

The data was labelled for all six cases by two experts. They labelled for image alone, text alone, and for the combination of image and text. They individually performed labelling for all 2100 data points. For the reliability of inter-rater agreement, we used Cohens Kappa (K) statistic measures (Berry and Mielke Jr, 1988). In our case we got K value to be 0.86 i.e., nearly perfect agreement. The details of our dataset can be seen in Table 2.

Fig. 4 Cyberbullying Post with having both image and comment are non-bullying

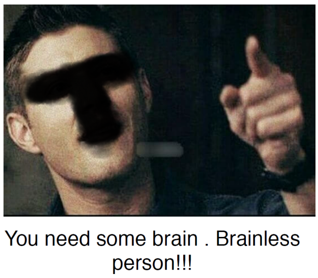

Fig. 5 Cyberbullying Post with having non-bullying image and bullying comment

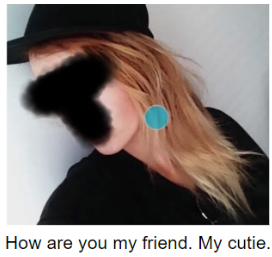

Fig. 6 Non-Cyberbullying Post with having both image and comment are non-bullying

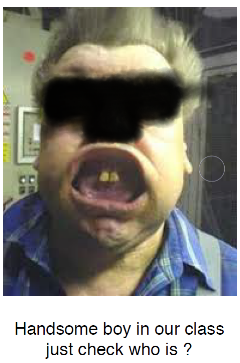

Fig. 7 Cyberbullying Post with having bullying image and non-bullying comment

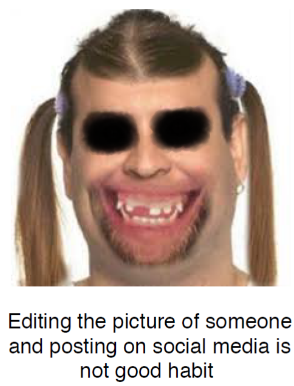

Fig. 8 Non-cyberbullying Post with having bullying image and non-bullying comment

Table 2 Description of dataset

Class	Image-label	Comment-label	Combined-label
Bullying	464	884	1481
Non-bullying	1636	1216	619
Total	2100	2100	2100

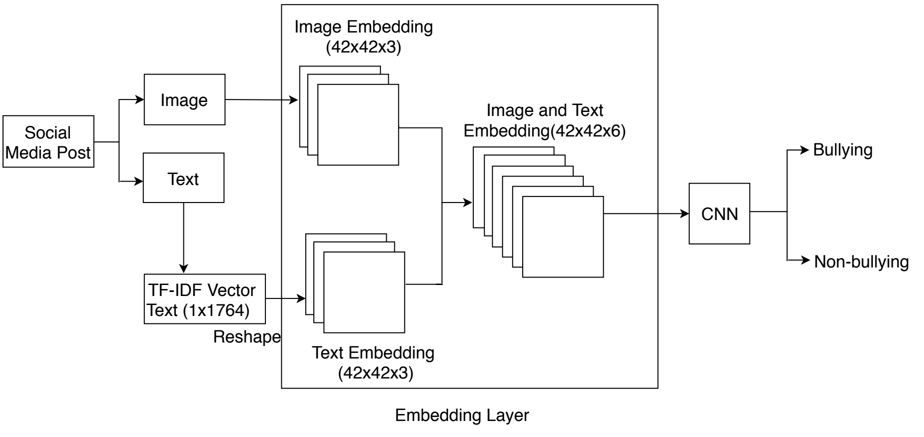

Fig. 9 Overview of the proposed approach

3.2 Overview of Proposed Approach

The proposed Convolutional Neural Network (CNN) based multi-modal system learns the integrated representation of post (containing image and text) to classify each post into bullying or non-bullying. As shown in Figure 9, the proposed model mainly contains four phases: (i) input preparation, (ii) embedding layer, (iii) convolutional layer, and (iv) output layer. The functioning of each phase are explained below:

3.2.1 Input Preparation

The input to the proposed system was integrated representation of image and text of the social media post. The processing of image and text files were done individually. It is easy to convert an image into the matrix because the image is made up of pixels. Therefore, after reading each image we got a matrix of $M \times N \times 3$ size, where M and N refer to the width and height of an image and 3 refers to the number of channels in a color image respectively. Each image in our dataset was a colored image therefore, it was represented in three channels red, green and blue. In contrast to the image, the processing of text was a little complex. To represent the text into a vector, we first created a bag of words of our dataset. There were 1802 unique words in our vocabulary. Then for each document in our dataset, we created a Term Frequency-Inverse Document Frequency (TF-IDF) vector representation.

3.2.2 Embedding Layer

In CNN, convolution on image and text follow different architecture. The convolution on the image is done in two dimensions whereas, the convolution on text is in one dimensional. Therefore, it was needed to give a unique representation of both text and image to the model so that convolution can be done. We had two options here, to either convert a two-dimensional image matrix into a single-dimensional vector or convert a one-dimensional word vector (TF-IDF) into the two-dimensional vector. Although we tried both cases, the results we got better for the latter case.

To convert a one-dimensional TF-IDF document vector, in a two-dimensional square matrix, we extracted the maximum size of the square we could form from 1802 words vocabulary. We found that the maximum word matrix we could form from 1802 unique words is 42×42 ($M \times N$), where the width and height of the matrix are 42 because next integer square matrix size (43×43) require 1849 words. Therefore we selected only 1764 top words for each document present in our dataset. We then converted each one dimensional document vector (with size 1×1764) into two-dimensional $42 \times 42 \times 1$ ($M \times N \times 1$) size document matrix, where 1 is the number of channel of the formed document matrix. To make a similar representation like image, each document matrix was replicated thrice and stacked one after another. After stacking, the final size of each document matrix become $42 \times 42 \times 3$, which was similar to the image matrix size of $M \times N \times 3$. But, in each image, M and N values were different. To make it uniform to final text document matrix, we padded each image into a $42 \times 42 \times 3$ size matrix. Finally, two $42 \times 42 \times 3$ matrices of image and text were embedded together to form a single $42 \times 42 \times 6$ matrix.

3.2.3 Convolution Layer

The embedded matrix of image and text ($42 \times 42 \times 6$) was then given as an input to the convolution layer for extracting combined features from it. For getting combined features set of image and text, we applied convolution operations on the embedding layer. Particularly, we applied three convolution layers with a filter size of 3×3 and ReLU activation function. The number of filters on first, second, and third convolution layers are 256, 256, and 128 respectively. To filtered out crucial features, we applied a max-pooling layer in a window of 2×2 after each convolution layer, to extract maximum features out of it.

3.2.4 Output Layer

The features extracted from the convolution layer were flattened into the one-dimensional vector, and then passed through two fully connected layers having 256 and 2 neurons. The output of the last fully connected layer was passed through the sigmoid activation function, that returns probabilities of a post of being in class bully and non-bully. Out of that, whichever probability is higher that represents the final class label of the post. To minimize loss while training the model, a binary cross-entropy loss function was used at the output layer. The hyper-parameters of the model used during our experiments are listed in Table 3.

Table 3 Hyper-parameters setting for the proposed multi-modal approach

Description	Values
Image size	42×42
Filter size	3×3
Number of Filters	256, 256, 128
Pooling size	2×2
Activation function	ReLU, Sigmoid
Dropout rate	0.5
Learning rate	0.001
Batch size	5
Loss Function	binary cross-entropy
Optimizer	Rmsprop
Epoch	1000

4 Results

This section discusses various results obtained while classifying the post into bullying and non-bullying classes. The proposed multi-modal approach based on Convolutional Neural Network (CNN) takes the post with text and image together as input, prepares its combined embedding, extracts combined features set, and classify it. In our dataset, we had both text and image, therefore the first challenge was to build a combined representation (or embedding) of text and image that could be given as an input to the proposed multi-modal system. It is worth mentioning that we have only two input representation that a CNN model accepts. That is we can either represent the data in one-dimensional (like we do in text data) or represent the data in two-dimensional (like we do in image data). Our target was to embed both image and text into a single representation either by one-dimensional (1-D) or two-dimensional (2-D) representation. For this, we did experimentation separately to evaluate the best representation between 1-D representation or 2-D representation. In the 1st attempt, we created a one-dimensional representation of both text and image. The combined one-dimensional representation is named as 1-D representation. In the 2nd attempt, we created two-dimensional representation for both text and image. They were combined to form the 2-D representation of input data. Our dataset had 619 and 1481 samples of non-bullying and bullying post respectively. To balance the dataset, we randomly picked 619 samples from 1481 samples of bullying posts. Finally, we got a balanced dataset having an equal number of samples of both (bullying and non-bullying) class. For the training, we took 75% samples and rest have been used for testing. We used Precision, Recall, and F1-score as performance metrics to evaluate the model. We discuss the different cases of our experiments in the following subsections.

4.1 1-D Representation

To process the data, at first, we embedded the text into 100-dimensional vectors using the TF-IDF vectorization. The maximum length to the comment was fixed to 25 for ease of experimentation. So, for each comment embedding dimension was 25×100. To make a similar representation of the image, we converted RGB-image to grayscale image and then converted the size of each image into 25×100. Finally, we stacked both image (25×100) and text (25×100) and got an integrated

Table 4 Results of classification using 1-D representation

Class	Results Precision	Recall	F1-score
Non-bullying	0.00	0.00	0.00
Bullying	0.54	1.00	0.70
Weighted average	**0.29**	**0.54**	**0.37**

Table 5 Results of classification using 2-D representation

Approach	n-gram	Precision	Recall	F1-score
$M \times N \times 4$	1	0.62	0.62	0.62
	2	0.57	0.57	0.57
	3	0.56	0.57	0.56
	1,2,3	0.60	0.57	0.51
$M \times N \times 6$	1	**0.63**	**0.63**	**0.63**
	2	0.60	0.60	0.60
	3	0.57	0.57	0.56
	1,2,3	0.58	0.57	0.57

representation of input in 50×100 dimension. The combined embedding was used as an input to the proposed model. Table 4 shows the performance of the model in the current setting. From Table 4, we can observe that the performance of the model was very poor because the model is not predicting anything for non-bullying class. So, we conclude that CNN could not extract the relevant features from combined representation when the input is in one-dimensional.

4.2 2-D representation

The other way to create combined representation was to represent text and image in the two-dimensional matrix, such that 2-D convolution could be applied to that. Image data is usually represented in a 2-D matrix, whereas the text data is in 1-D vector. To convert text data from 1-D to 2-D, we created a TF-IDF vector of 1802 words (our vocabulary size). Then we created the combined representation for two cases. First, we integrated 3 channel of RGB image and 1 channel of text into $M \times N \times 4$ size matrix, and in second case we integrated 3 channel of both image and text into $M \times N \times 6$ size matrix (As explained in Methodology section the maximum size of the square matrix can be formed as (42×42). In the first case, we stacked each image of size $42 \times 42 \times 3$ and $42 \times 42 \times 1$ size of text. Finally, we got together (image and text) a single $42 \times 42 \times 4$ matrix. Next, we tried to make a similar representation like image, each document matrix was replicated thrice, and kept one after another. Finally, we got together $M \times N \times 6$ matrix. We performed experiments for both cases with different n-grams features like 1-gram, 2-gram, 3-gram, and together with 1,2,3-grams. Where n-grams represent the number of words in a sequence. Table 5 shows the results (in a weighted average of non-bullying and bullying classes) of our experiments using three convolutional layers (3-CNN) with the filter size of 256, 256, 128 for first, second, and third layers respectively. We got the best result for $M \times N \times 6$ matrix with 1-gram features. So, in our further experiment, we stuck with $M \times N \times 6$ matrix and 1-gram features.

Table 6 Results of classification varying with convolutional layers and number of filters

Approach	Filter size	Class	Precision	Recall	F1-score
1-CNN	256	Non-bullying	0.64	0.68	0.66
		Bullying	0.71	0.66	0.68
		Weighted average	0.67	0.67	0.67
	512	Non-bullying	0.66	0.63	0.65
		Bullying	0.69	0.72	0.70
		Weighted average	0.68	0.68	0.68
	1024	Non-bullying	0.65	0.72	0.68
		Bullying	0.73	0.66	0.70
		Weighted average	0.69	0.69	0.69
	2048	Non-bullying	0.67	0.61	0.64
		Bullying	0.69	**0.74**	0.71
		Weighted average	**0.68**	**0.68**	**0.68**
2-CNN	256, 128	Non-bullying	0.61	0.74	0.67
		Bullying	0.72	0.59	0.65
		Weighted average	0.67	0.66	0.66
	1024, 512	Non-bullying	0.65	0.69	0.67
		Bullying	0.72	0.67	0.69
		Weighted average	0.68	0.68	0.68
3-CNN	1024, 512, 256	Non-bullying	0.59	0.47	0.52
		Bullying	0.61	0.71	0.66
		Weighted average	0.60	0.60	0.60

4.3 Experimenting with different Convolution layers with different filter sizes

Next, we needed to determine how many layers of convolution is better for our task. So, we experimented with one, two, and three convolution layers. We also needed to determine the most appropriate size of filters to be used in convolution layers. Table 6 shows the different combinations of convolution layers with a different combination of filters. The number of filters used in each layer of convolution was mentioned in Table 6. Our main target was to identify bullying cases more accurately. So, our main performance matrix to compare the best model is the recall value of the bullying class. We got the best result for one convolutional (1-CNN) layer with a filter size of 2048. The best result is shown in bold in Table 6. Out of actual bullying cases, in 74% of the cases, we got the correct prediction. The confusion matrix of the best performing proposed model is shown in Figure 10.

5 Discussion

One of the main findings of this research is that the combined embedding of text and image performed better in 2-D representation in comparison of 1-D representation. Combining features from a 1-D image vector lost the image characteristics due to the pixel configuration of an image. However, the convolution captures the rationale of the 2-D image better. That is nearer the pixels, more closer the relationship with each other. Therefore, the convolution layer always preserves the 2-D spacial information. But, 1-D representation destroyed it when they were given input to CNN. With this 2-D representation, the proposed multi-modal approach can currently predict the 74% of bullying posts out of 100 true cases of bullying posts as shown in Table 6. Based on the results of the present study, we can deduce

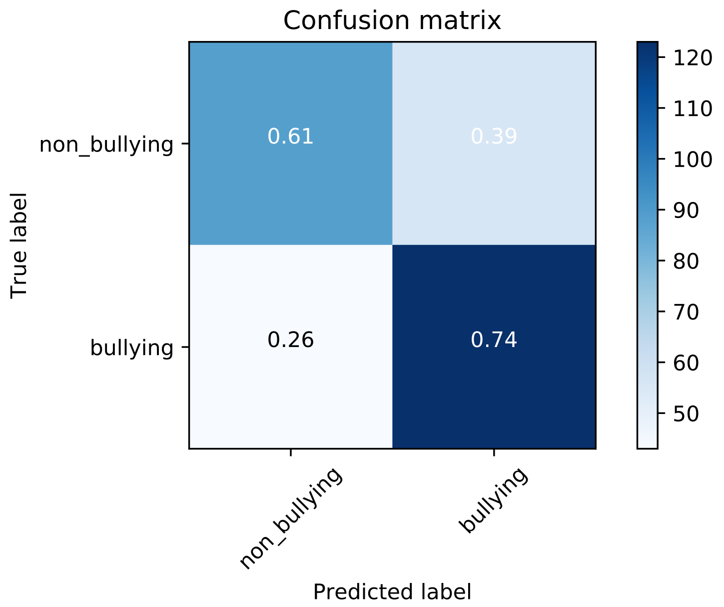

Fig. 10 Confusion matrix of best result of proposed approach

that the 1-gram TF-IDF features when replicated three times and stacked with three channels of image and given as a 1-CNN, 2-Dense with a filter size of 2048 and dropout of 0.5 better model compared to others.

While creating a unified 2-D representation of text and image, we followed two approaches. In the first approach, we combined $M \times N \times 3$ image with $M \times N \times 1$ text matrix, which results in $M \times N \times 4$ combined representation. Whereas, in the second approach we combined $M \times N \times 3$ image with $M \times N \times 3$ text matrix to give a combined representation of $M \times N \times 6$. In Cyberbullying identification task, we found that $M \times N \times 6$ representation is better than $M \times N \times 4$ which can be observed from Table 5. The reason behind the performance of $M \times N \times 6$ being better than $M \times N \times 4$ model is in $M \times N \times 6$ setting, the weight of text is 3 times more than $M \times N \times 4$ setting. Whereas the same image can have a different sense but in the text generally has a clear sense. Therefore, the text has a clearer meaning than the image. Our results deduce that in Cyberbullying identification task we should give more weight to text than image.

Our next finding is 1-gram TF-IDF features are better than 2-gram, 3-gram, or together 1,2,3-gram TF-IDF features for combining image & text through 2-D representation which can be observed from Table 5. To convert text of $M \times N \times 1$ matrix into $M \times N \times 3$ matrix, we used n-gram approach. We tried several settings with 1-gram, 2-gram, 3-gram or together 1,2,3-gram TF-IDF features. However, it is found that the system performed best when 1-gram TF-IDF features were replicated thrice to convert $M \times N \times 1$ representation of text. It was combined with $M \times N \times 3$ image matrix to give a unified representation of image and text in $M \times N \times 6$. As we know that n-gram features extract only the most important pieces of information from long text strings. The reason behind 1-gram is performing better than 2-gram, 3-gram, and together 1,2,3-gram TF-IDF features is that two-dimensional convolution operation on the individual word is more meaningful than a collection of words together which is used in 2-gram, 3-gram, together 1,2,3-gram features.

Our last finding is that 1-CNN is performing better than 2-CNN, and 3-CNN which can be observed from Table 6. The reason behind single layer CNN is performing better than multiple layers of CNN is in multiple layers of CNN millions of weights are used to model a system, and this usually causes overfitting. Therefore, in the combined representation of image and text, the simple model is performing better than the complex model.

It should be noted that the current approach can classify the bullying post with good performance measures. We have considered heterogeneous data (image and text) to train the single model for both data. Our simple one convolutional layer is performing better than multiple layers of convolutional layers which is a more complex model. The best results are shown in Figure 10, the observation from these results is that we got 74% and 61% recall values for bullying and non-bullying class respectively. Overall, we got the best results with F1-score of 71% and 64% for bullying and non-bullying class respectively can be seen in Table 6. Our observation is that the simple model is better than the complex model when multi-modal data are embedded properly.

5.1 Theoretical and Practical Implication

The current research expands the prosperous literature on the identification of Cyberbullying by proposing a novel unified multi-modal approach. The main theoretical implication of this work is the integration of image and text into a single representation. That means we should not require two parallel systems to process heterogeneous data (image and text). The single system works for both type of data and learns the structure of image and text from single representation. Irrespective of the Cyberbullying task, the proposed system can also be utilized in the case of disaster management, emotion detection and many other cases where both image and text are a major source of information. We hope our work can be a benchmark for combining multi-modal data representation.

The major practical implication of this work is it can be a better tool for the identification of heterogeneous social media posts where the post has a different form of data. The present system can be installed on top of any classification task which can be benefited from these settings. This will help online users to use social media as a safe environment to interact with other online users in the smart city.

6 Conclusions and Future work

Social mining is generally understood as representing, analyzing and extracting enforceable trends and patterns from raw data on social media. The current research aimed at combining both visual and textual characteristics to identify bullying posts on social media. This paper has introduced a novel framework to identify Cyberbullying instances with the new integrated representation of image and text. This important contribution provides an analytical background that opens the way to combine different forms of data to be trained in a single system instead of parallel systems where different systems were used for different types of data. Our proposed system can correctly identify 74% of cases of bullying class. Overall our system got 68% weighted average F1-score of both (bullying and non-bullying) classes. We found that a single layer of convolution with a larger filter size is better than multiple layers of convolution with a lesser number of filters.

We have only considered the image and text for Cyberbullying detection task but audio, video, URLs of the post also can be useful information of consideration for identifying bullying scenarios. Finally, despite introducing a unified representation of different modalities, future research should aim to determine the proper weight of text and image into a Cyberbullying identification task.

Acknowledgements

Kirt Kumari would like to acknowledge the Ministry of Electronics and Information Technology (MeitY), Government of India for the financial support to her during the research work through Visvesvaraya Ph.D. Scheme for Electronics and IT.

Compliance with Ethical Standards

Funding: This study was funded by the Ministry of Electronics and Information Technology (MeitY), Government of India through Visvesvaraya PhD Scheme for Electronics and IT.
Conflict of Interest: The first author has received research grants from the Ministry of Electronics and Information Technology (MeitY), Government of India through Visvesvaraya PhD Scheme for Electronics and IT. Authors, Jyoti Prakash Singh, Yogesh Kumar Dwivedi, and Nripendra Pratap Rana declare that they have no conflict of interest.
Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

Al-garadi, M. A., K. D. Varathan, and S. D. Ravana (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior 63*, 433–443.

Ali, A. M. and P. Angelov (2018). Anomalous behaviour detection based on heterogeneous data and data fusion. *Soft Computing 22*(10), 3187–3201.

Alkhammash, E. H., J. Jussila, M. D. Lytras, and A. Visvizi (2019). Annotation of smart cities twitter micro-contents for enhanced citizens engagement. *IEEE Access 7*, 116267–116276.

Badjatiya, P., S. Gupta, M. Gupta, and V. Varma (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760. International World Wide Web Conferences Steering Committee.

Berry, K. J. and P. W. Mielke Jr (1988). A generalization of cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement 48*(4), 921–933.

Bhat, C. S., M. A. Ragan, P. R. Selvaraj, and B. J. Shultz (2017). Online bullying among high-school students in india. *International Journal for the Advancement of Counselling 39*(2), 112–124.

Bohra, A., D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava (2018). A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of Peoples Opinions, Personality, and Emotions in Social Media*, pp. 36–41.

Burnap, P. and M. L. Williams (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet 7*(2), 223–242.

Chatzakou, D., N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali (2017). Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pp. 13–22. ACM.

Chen, H., S. Mckeever, and S. J. Delany (2017). Harnessing the power of text mining for the detection of abusive content in social media. In *Advances in Computational Intelligence Systems*, pp. 187–205. Springer.

Chen, Y., Y. Zhou, S. Zhu, and H. Xu (2012). Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pp. 71–80. IEEE.

Chui, K., M. Lytras, and A. Visvizi (2018). Energy sustainability in smart cities: Artificial intelligence, smart monitoring, and optimization of energy consumption. *Energies 11*(11), 2869.

Dadvar, M. and F. De Jong (2012). Cyberbullying detection: a step toward a safer internet yard. In *Proceedings of the 21st International Conference on World Wide Web*, pp. 121–126. ACM.

Dadvar, M., D. Trieschnigg, and F. de Jong (2014). Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Canadian Conference on Artificial Intelligence*, pp. 275–281. Springer.

Davidson, T., D. Warmsley, M. Macy, and I. Weber (2017). Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Dinakar, K., B. Jones, C. Havasi, H. Lieberman, and R. Picard (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS) 2*(3), 18.

Hosseinmardi, H., S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishr (2015). Prediction of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1508.06257*.

Hosseinmardi, H., R. I. Rafiq, R. Han, Q. Lv, and S. Mishra (2016). Prediction of cyberbullying incidents in a media-based social network. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 186–192. IEEE.

Jones, L. M., K. J. Mitchell, and D. Finkelhor (2013). Online harassment in context: Trends from three youth internet safety surveys (2000, 2005, 2010). *Psychology of violence 3*(1), 53.

Kornblum, J. (2008). Cyberbullying grows bigger and meaner with photos, video. *USA Today*.

Kumari, K., J. P. Singh, Y. K. Dwivedi, and N. P. Rana (2019). Aggressive social media post detection system containing symbolic images. In *Conference on e-Business, e-Services and e-Society*, pp. 415–424. Springer.

League, A. D. (2011). Glossary of cyberbullying terms. adl. org.

Lytras, M. and A. Visvizi (2018). Who uses smart city services and what to make of it: Toward interdisciplinary smart cities research. *Sustainability 10*(6), 1998.

Lytras, M., A. Visvizi, L. Daniela, A. Sarirete, and P. Ordonez De Pablos (2018). Social networks research for sustainable smart education. *Sustainability 10*(9), 2974.

Nahar, V., X. Li, and C. Pang (2013). An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering 3*(5), 238.

Paavola, J., T. Helo, H. Jalonen, M. Sartonen, and A. Huhtinen (2016). Understanding the trolling phenomenon: The automated detection of bots and cyborgs in the social media. *Journal of Information Warfare 15*(4), 100–111.

Pater, J. A., A. D. Miller, and E. D. Mynatt (2015). This digital life: A neighborhood-based study of adolescents' lives online. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 2305–2314. ACM.

Reynolds, K., A. Kontostathis, and L. Edwards (2011). Using machine learning to detect cyberbullying. In *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on*, Volume 2, pp. 241–244. IEEE.

Salawu, S., Y. He, and J. Lumsden (2017). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing* (1), 1–1.

Seiler, S. J. and J. N. Navarro (2014). Bullying on the pixel playground: Investigating risk factors of cyberbullying at the intersection of childrens online-offline social lives. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace 8*(4).

Singh, V. K., S. Ghosh, and C. Jose (2017). Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2090–2099. ACM.

Steiner-Correa, F., M. I. Viedma-del Jesus, and A. Lopez-Herrera (2018). A survey of multilingual human-tagged short message datasets for sentiment analysis tasks. *Soft Computing 22*(24), 8227–8242.

Sui, X., Z. Chen, L. Guo, K. Wu, J. Ma, and G. Wang (2017). Social media as sensor in real world: movement trajectory detection with microblog. *Soft Computing 21*(3), 765–779.

Tommasel, A., J. M. Rodriguez, and D. Godoy (2018). Textual aggression detection through deep learning. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 177–187.

Torres-Ruiz, M. J. and M. D. Lytras (2016). Urban computing and smart cities applications for the knowledge society. *International Journal of Knowledge Society Research (IJKSR) 7*(1), 113–119.

Van Royen, K., K. Poels, W. Daelemans, and H. Vandebosch (2015). Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. *Telematics and Informatics 32*(1), 89–97.

Visvizi, A., J. Jussila, M. D. Lytras, and M. Ijäs (2019). Tweeting and mining oecd-related microcontent in the post-truth era: A cloud-based app. *Computers in Human Behavior*.

Visvizi, A. and M. D. Lytras (2018). Rescaling and refocusing smart cities research: From mega cities to smart villages. *Journal of Science and Technology Policy Management 9*(2), 134–145.

Visvizi, A., M. D. Lytras, E. Damiani, and H. Mathkour (2018). Policy making for smart cities: Innovation and social inclusive economic growth for sustainability. *Journal of Science and Technology Policy Management 9*(2), 126–133.

Wang, L., J. Zhang, P. Liu, K.-K. R. Choo, and F. Huang (2017). Spectral–spatial multi-feature-based deep learning for hyperspectral remote sensing image classification. *Soft Computing 21*(1), 213–221.

Waseem, Z. and D. Hovy (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pp. 88–93.

Yin, D., Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards (2009). Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB 2*, 1–7.

Zhao, R. and K. Mao (2016). Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing, vol. PP* (99), 1–1.

Publishers Note